The Fourth international Conference Theoretical and Applied Aspects of Program Systems Development
(Ukraine, Berdyansk, 4-9 September 2007)

ОЛЕЦЬКИЙ О.В.,
факультет інформатики Національного Університету "Києво-Могилянська Академія",
кандидат технічних наук, доцент
E-mail: oletsky@ukma.kiev.ua

ДО ПРОБЛЕМИ ОНТОЛОГІЧНО-ОРІЄНТОВАНОГО ПОШУКУ В ІНФОРМАЦІЙНИХ СИСТЕМАХ

Сьогодні стрімко розвиваються засоби, які дозволяють пошуковій системі знайти серед великої множини документів ті, які потрібні користувачеві. При цьому ключову роль відіграють методи оцінки релевантності (відповідності) документів запитові. В ідеалі ці оцінки повинні, з одного боку, бути достатньою мірою математично обґрунтованими, аз іншого — відповідати інтуїтивним очікуванням. При оцінці релевантності необхідно також враховувати мету пошуку, персональні характеристики користувача, історію його попередніх пошуків тощо.

Сьогодні існує велика кількість методів отримання таких оцінок [1-3 та ін.]. Проте тут домінують евристичні підходи, і навіть саме поняття "відповідність запитові користувача" є недостатньо формалізованим і носить інтуїтивний та суб'єктивний характер. Крім того, очевидною є необхідність розвивати семантично-орієнтовані методи пошуку, які б враховували не тільки власне тексти документів та пов'язані з ними метадані, а максимально спиралися б на семантику предметної області. В цьому контексті важливою є побудова цілісних формальних моделей — з одного боку, достатньо простих, а з іншого — таких, які б дозволяли розв'язувати достатньо складні задачі. Крім того, подібні формалізації повинні брати до уваги існуючі рекомендації та технології з огляду на логіку і перспективи їх розвитку.

Як основа для побудови таких моделей може розглядатися підхід на основі онтологічно-документного моделювання. В основі моделі інформаційної бази веб-орієнтованої системи повинні лежати дві вузлові компоненти: онтологія предметної області та множина документів, а також зв'язки між цими компонентами. Таким чином, необхідно здійснювати аналіз графу, який складається не з логічно розрізнених документів, що є типовим для більшості сучасних веб-орієнтованих систем, а з описів реальних класів предметної області, їх екземплярів та зв'язків між ними, а також пов'язаних з ними документів. Таким чином, слід говорити про "занурення" множини документів, які можуть бути статичними або генеруватися динамічно, в загальну семантику предметної об-

ласті. Онтології, які беруться до уваги, можуть бути багаторівневими — від загальних онтологій інформаційних ресурсів (найбільш відома з них — Dublin Core) до предметних онтологій, для яких повинна бути забезпечена можливість розвитку і поповнення.

У цьому контексті слід також зазначити, що проблема підвищення повноти та релевантності інформаційного пошуку найтіснішим чином пов'язана з проблемою семантично-орієнтованого управління вмістом веб-сайту та навігації по ньому [4, 5] (прототип такої системи на основі Protege та pOWL описано в [5]) . Дійсно, незалежно від того, чи користувач вийшов на певний вузол онтології в процесі навігації по сайту, чи він ввів відповідне ключове слово в полі введення — мова йде про формування переліку документів, пов'язаних з даним вузлом, і ранжування цих документів за мірою релевантності. Таким чином може вирішуватися проблема інтеграції методів семантично-орієнтованого управління контентом, зокрема семантично-орієнтованої навігації, з одного боку, та семантичного пошуку — з іншого боку.

Більш формально, підхід на основі онтологічно-документного моделювання ґрунтується на формалізації зв'язків між семантикою предметної області та множиною документів; для такої формалізації природно використати формальні моделі онтологій [6-8]. В [4, 5] зв'язки між онтологією предметної області W та множиною документів D описуються наступним чином.

Якщо онтологія розглядається як трійка <Q, R, F>, де Q — множина класів, які відповідають поняттям предметної області, R — множина зв'язків між ними, а F — множина функцій інтерпретації, то розширена онтологія описується як трійка <Q*, R*, F*>, де Q* — множина класів разом з їх екземплярами, R*- множина зв'язків між цими елементами, а F* — множина функцій інтерпретації, визначених у найпростішому випадку на елементах з Q*, R* та Q* x R* x F*. Тоді елементи D можуть бути значеннями функцій з F*. Іншими словами, будемо вважати документ d релевантним відносно W*, якщо існують хоча б один вузол w та функція інтерпретації f, такі що d=f(w).

Наведене співвідношення може стати основою для динамічного формування переліку споріднених документів. Для документа d=f(w) спорідненими документами можуть вважатися, зокрема, такі:

- всі документи s такі, що s=h(w), h€F * (тобто документи, пов'язані з тим самим вузлом іншими функціями інтерпретації);
- всі документи s такі, що s=g(u), де функції інтерпретації g пов'язані з f, вузли u пов'язані з w.

Безумовно, можуть задаватися і інші правила, що визначають спорідненість документів.

На основі цього можна оцінювати міри релевантності документів та міри спорідненості між ними. Очевидно, поняття релевантності документа запитові користувача та спорідненості документів між собою тісно пов'язані: можна вважати, що споріднені документи будуть мати близькі міри релевантності, якщо специфіка конкретної задачі не дозволяє встановити інше.

Тепер проблема полягає в тому, щоб знайти документи, які прямо пов'язані з запитами, а також споріднені з ними. Перш за все, запит співставляється з онтологією предметної області, знаходяться відповідні вузли та пов'язані з ними документи. Далі можна використати метод поширення активації, застосування якого до задачі інформаційного пошуку було описано в [9]. Ініціюється певний хвильовий процес пошуку, під час якого активація вузлів може здійснюватися в двох просторах:

- пошук в просторі документів, при цьому міри спорідненості мо жуть задаватися явним чином або визначатися динамічно; наприк лад, на основі просторово-векторної моделі [1];
- пошук на графі, який задає розширену онтологію.

При цьому розрахунок мір релевантності здійснюється на основі коефіцієнтів, що пов'язуються з окремими типами зв'язків між вузлами розширеної онтології, а також з окремими типами функцій інтерпретацій. В результаті будуть сформовані множини понять і документів, в тій чи іншій мірі релевантних запиту. В типовому випадку кожний елемент цих множин буде поданий у вигляді (u, $m_1(u)$, ..., $m_r(u)$), де u — знайдений вузол, $m_i(u)$ — міра релевантності цього вузла, обчислена за і-м критерієм, можливо, недостовірна або нечітка. Припускається навіть динамічне породження нових критеріїв, якщо задати процедуру такого породження.

Основна проблема, яка виникає при цьому, пов'язана зі значними обсягами інформації, і відповідно — зі значною часовою складністю. Тому ключовим є наступне питання: як спрямувати процес поширення активації в потрібному напрямку, і яким повинен бути критерій зупинки цього процесу? Важливою є також проблема комбінування критеріїв, яка полягає в тому, щоб перейти від кількох мір релевантності документа за різними критеріями до однієї комбінованої міри релевантності.

Необхідно також відійти від явного задания коефіцієнтів зв'язків — ці коефіцієнти повинні налаштовуватися динамічно. Інтелектуальна пошукова система повинна також мати в своєму розпорядженні засоби управління експериментом, які дозволяли б здійснювати порівняльний

аналіз різних алгоритмів інформаційного пошуку за тими чи іншими критеріями та експериментальний підбір параметрів цих алгоритмів, інші параметри хвильового процесу поширення активації, а також комбінувати різні методи пошуку.

Для розв'язання перелічених проблем видається доцільним застосовувати методики випадкового керування інформаційним пошуком [10]; зокрема генетичні алгоритми, які добре зарекомендували себе для розв'язання ряду перебірних задач [11, 12]. В контексті, який розглядається, можна виділити як мінімум два аспекти застосування цих алгоритмів:

- власне для вибору найбільш перспективної підмножини документів, серед яких документи, потрібні користувачеві, будуть міститися з максимальною ймовірністю;
- для експериментального підбору параметрів хвильового процесу поширення активації.

Описану методику найбільш зручно використовувати для локального пошуку, і особливо — для проблемно-орієнтованих інформаційних просторів (порталів знань) [3, 13], для яких характерною є тематична однорідність та достатньо висока зв'язність інформаційних ресурсів. Зокрема, на такі портали знань можуть бути перетворені цифрові бібліотеки.

Для класичних пошукових систем, які зберігають глобальну або регіональну базу пошукових образів документів, характерними є значно більші обсяги інформації. Крім того, онтологічно-орієнтовані глобальні пошукові системи повинні мати в своєму складі засоби для роботи з онтологіями різних предметних областей, а також механізм, який дозволяє підключати ті чи інші засоби в залежності від запиту користувача. Може виникнути ситуація, коли доводиться працювати з кількома онтологіями, які стосуються однієї предметної області, і тоді виникає проблема їх інтеграції.

ЛІТЕРАТУРА

1. Сэлтон Дж. Автоматическая обработка, хранение и поиск инфор мации. — М.: Сов.радио, 1973. — 560 с.
2. Ландэ Д.В. Поиск знаний в Интернет. — М: Изд. дом "Вильяме", 2005. —272 с.
3. Гриценко В.И., Духновская К.К., Урсатьев А.А. Поисковый сервис. Проблемы, технологии, перспективы // УСиМ, 2006, №2. — С. 81-92.
4. Олецький О.В. Застосування формальних моделей онтологій для формалізації інформаційних потоків у системах управління кон-

тентом. //Теоретичні та прикладні аспекти побудови програмних систем. Матеріали міжнародної конференції TAAPSD'2005, Київ, 7-9 грудня 2005 р. — С. 26-29.

5. Діренко І.С., Олецький О.В. Система управління вмістом вцб-ресурсів на основі онтологічно-документного моделювання //Теоре тичні та прикладні аспекти побудови програмних систем. Матеріа ли міжнародної конференції TAAPSD'2006,. Київ, грудень 2006 р. — С.171-176.
6. Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем. — СПб: Питер, 2000. — 384 с.
7. Плескач В.Л., Рогушина Ю.В. Агентні технології. — К.: Київ. нац. торг.-екон. ун-т, 2005. — 338 с.
8. Проскудіна Г.Ю., Овдій О.М. Онтології в інформаційних системах. //Теоретичні та прикладні аспекти побудови програмних систем. Спеціальний випуск Вісника Київського Нсіціонального університе ту ім.Т.Г.Шевченка, 2004. — С. 164-169.
9. Глибовець М.М. Моделі та методи створення і супроводу високоп родуктивного розподіленого навчального середовища. Авторефе рат дисертації на здобутгя наукового ступеня доктора фізико-математичних наук. — Національний університет "Києво-Могилянська Академія", Київ, 2006.
10. Глибовець М.М., Олецький О.В. Про деякі підходи до проблеми ін формаційного керування випадковим пошуком. //Dynamical System Modelling and Stability Investigation. Thesis of Conference Reports, May 22-25, 2007.—-С.370:
11. Глибовец Н.Н., Медведь С.А. Генетические алгоритмы и их использование для решения задач составления раслисанияУ/Кибернетика и системный анализ, 2003, №1. — С.95-108.
12. Рутковская Д., Пилиньский М., Рутковский Л. Нейронные сети, генетические алгоритмы и нечеткая логика. — М.: Горячая линия — Телеком, 2004. — 452 с.
13. Боровикова О.И., Загорулько Ю.А. Организация порталов знаний на основе онтологии //Тр. Междунар. сем. Диалог'2002 "Компью терная лингвистика и интеллектуальные технологии" — Т.2. — Протвино, 2002. — С.76 — 82.